A TEXT-TO-SPEECH CONVERSION SYSTEM FOR INTERLOCKING WITH MULTIMEDIA AND A METHOD FOR ORGANIZING INPUT DATA OF THE SAME

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention relates to a text-to-speech conversion system (hereinafter, referred to as TTS) for interlocking with multimedia and a method for organizing input data of the same, and more particularly to a text-to-speech conversion system (TTS) for interlocking with multimedia and a method for organizing input data of the same for enhancing the natural of synthesized speech and accomplishing synchronization between multimedia and TTS by defining additional prosody information, the information required to interlock TTS with multimedia, and interface between these information and TTS for use in the production of the synthesized speech.

Description of the Related Art

Generally, the function of the speech synthesizer is to provide different forms of information for a man using a computer. To this end, the speech synthesizer should serve the user with synthesized speech with high quality from a given text. In addition, for the interlock with database produced in multimedia environment such as moving picture or animation, or

a variety of media provided from a counterpart of conversion, the speech synthesizer should produce the synthesized speech to be synchronized with theses media. Particularly, the synchronization of TTS with multimedia is essential to provide the user with service with high quality.

As shown in Fig. 1, typically, a conventional TTS goes through the process consisting of 3 steps as follows until the synthesized speech is produced from on inputted text.

In a first step, a language processor **1** converts the text into a series of phoneme, presumes prosody information and symbolizes this information. Symbol of prosody information is presumed from a boundary of the phrase and paragraph, a location of accent in word, a sentence pattern, and so on using the analysis result of syntax.

In a second step, a prosody processor **2** calculates a value of prosody control parameter from the symbolized prosody information using a rule and a table. Prosody control parameter includes duration of phoneme, pitch contour, energy contour, and pause interval information.

In a third step, a signal processor **3** produces a synthesized speech using a synthesis unit database **4** and the prosody control parameter. In other words, this means that the conventional TTS should presume the information associated with the natural and speech rate in the language processor **1** and the prosody processor

3

2 only by the inputted text.

Further, the conventional TTS has simple function to output data inputted by the unit of sentence as the synthesized speech. Accordingly, in order to output sentences stored in a file or sentences inputted through a communication network as the synthesized speech in succession, a main control program which reads sentences from the inputted data and transmits them to an input of TTS is required. Such a main control program includes a method to separate the text from the inputted data and then output the synthesized speech once from the beginning to the end, a method to produce the synthesized speech in interlock with a text editor, a method to look up the sentences by use of a graphic interface and produce the synthesized speech, and so on, but the object to which these methods are applicable is restricted to the text.

At present, studies on TTS have considerably advanced for the vernacular language in different countries and a commercial use has been accomplished in some countries. However, this is in situation of the only use for the syntheses of speech from the inputted text. In addition, by a prior organization, since it is impossible to presume from only the text the information required when moving picture is to be dubbed by use of TTS or when the natural interlock between the synthesized speech and multimedia such as animation is to be implemented, there is no method to realize these functions. Furthermore, there is also no result of the studies on use of additional data for

enhancement of the natural in the synthesized speech and organization of these data.

SUMMARY OF THE INVENTION

Therefore, it is an object of the present invention to provide a text-to-speech conversion system (TTS) for interlocking with multimedia and a method for organizing input data of the same for enhancing the natural of synthesized speech and accomplishing synchronization of multimedia with TTS by defining additional prosody information, the information required to interlock TTS with multimedia, and interface between these information and TTS for use in the production of the synthesized speech.

In order to accomplish the above object, a TTS for interlocking with multimedia according to the present invention comprises a multimedia information input unit for organizing text, prosody, the information on synchronization with moving picture, lip-shape, and the information such as individual property; a data distributor by each media for distributing the information of the multimedia information input unit into the information by each media; a language processor for converting the text distributed by the data distributor by each media into phoneme stream, presuming prosody information and symbolizing the information; a prosody processor for calculating a value of prosody control parameter from the symbolized prosody information using a rule and a table; a synchronization adjustor for

adjusting the duration of the phoneme using the synchronization information distributed by the data distributor by each media; a signal processor for producing a synthesized speech using the prosody control parameter and data in a synthesis unit database; and a picture output apparatus for outputting the picture information distributed by the data distributor by each media onto a screen.

In order to accomplish the above object, a method for organizing input data of a text-to-speech conversion system (TTS) for interlocking with multimedia comprises the steps of: classifying multimedia input information organized for enhancing the natural of synthesized speech and implementing the synchronization of multimedia with TTS into text, prosody, the information on synchronization with moving picture, lip-shape, and individual property information in a multimedia information input unit; distributing the information classified in the multimedia information input in a data distributor by each media, based on respective information; converting text distributed in the data distributor by each media into phoneme stream, presuming prosody information and symbolizing the information in a language processor; calculating a value of prosody control parameter other than prosody control parameter included in multimedia information in a prosody processor; adjusting the duration every each phoneme in a synchronization adjustor so that processing result in the prosody processor may be synchronized with a picture signal according to input of the synchronization information; producing the synchronized speech in a signal processor using the prosody

information from the data distributor by each media, the processing result in the synchronization adjustor, and a synthesis unit database; and outputting the picture information distributed by the data distributor by each media onto a screen in a picture output apparatus.

BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing and other objects, features, aspects of the present invention will become more apparent from the following detailed description of the present invention when taken in conjunction with the accompanying drawings.

FIG. 1 is a constructional view of a conventional text-to-speech conversion system.

FIG. 2 is a constructional view of a hardware to which the present invention is applied.

FIG. 3 is a constructional view of a text-to-speech conversion system according to the present invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

Now, the present invention will be described in detail by way of the preferred embodiment.

Referring to FIG. 2, a constructional view of hardware to which the present invention is applied is shown. In FIG. 2, the hardware consists of a multimedia data input unit **5**, a central

processing unit **6**, a synthesis database **7**, a digital to analog (D/A) converter **8**, and a picture output apparatus **9**.

The multimedia data input unit **5** is inputted with data composed of multimedia such as picture and text and outputs this data to the central processing unit **6**.

The central processing unit **6** distributes the multimedia data input of the present invention, adjusts synchronization, and performs algorithm based therein to produce synthesized speech.

The synthesis database **7** is a database used in the algorithm for producing the synthesized speech. This synthesis database **7** is stored in a storage device and transmits necessary data to the central processing unit **6**.

The digital to analog (D/A) converter **8** converts the synthesized digital data into analog signal and outputs the analog signal.

The picture output apparatus **9** outputs inputted picture information onto a screen.

Table 1 and 2 are algorithms illustrating the state of organized multimedia input information, which consists of text, prosody, the information on synchronization with moving picture, lip-shape, and individual property information.

Table 1

~~[Table 1].~~

Syntax

```
TTS_Sequence() {
TTS_Sequence_Start_Code
Prosody_Enable
Video_Enable
Lip_Shape_Enable
Start_Any_Place
do{
      TTS_Sentence()
      }while(next_bits()==TTS_Sentence_Start_Code
}
```

Here, the TTS_Sequence_Start_Code is a bit string represented with Hexadecimal 'XXXXX' and means a start of TTS sentence.

The TTS_Sentence_ID is a 10-bit ID and represents a proper number of each TTS data stream.

The language_Code represents an object language such as Korean language, English language, German language, Japanese language, French language etc,. to be synthesize.

The prosody_Enable is a 1-bit flag and has a value of '1' when a prosody data of original sound is included in an organized data.

The Video_Enable is a 1-bit flag and has a value of '1' when a TTS is interlocked with moving picture.

The Lip_Shape_Enable is a 1-bit flag and has a value of '1' when a lip-shape data is included in an organized data.

The Trick_Mode_Enable is a 1-bit flag and has a value of '1' when a data is organized to support a trick mode such as stop, restart, forward and backward.

[Table 2]

Syntax

TTS_Sentence(){
TTS_Sentence_Start_Code
Silence
if(Silence){
Silence_Duration
}
else{
Gender
Age
if(!Video_Enable){
Speech_Rate
}
Length_of_Text
TTS_Text
Position_in_Sentence
if(Prosody_Enable){
Number_of_phonemes
Dur_Enable
F0_Enable
Energy_Enable
for(j=0 ; j<Number_of_phonemes ; j++){
Symbol_each_phoneme
Dur_each_phoneme
F0_contour_each_phoneme
Energy_contour_each_phoneme
}
if(Video_Enable){
Sentence_Duration
Position_in_Sentence
offset
}
if(Lip_Shape_Enable){
Number_of_Lip_Event
for(j=0 ; j<Number_of_Lip_Event ; j++){
Lip_in_Sentence
Lip_Shape
}
}
}
}

Here, the TTS_Sentence_Start_Code is a bit string represented with Hexadecimal 'XXXXX' and means a start of TTS sentence. And the TTS_Sentence_Start_Code is a 10-bit ID and represents a proper number of each TTS data stream.

The TTS_Sentence_ID is a 10-bit ID and represents a proper number of each TTS sentence existed in the TTS stream.

The Silence become a '1' when a present input frame of 1-bit flag is silence speech section.

At stage of the Silence_Duration, a duration time of present silence speech section is represented by milliseconds.

At stage of the Gender, gender is distinguished from a synthesized speech.

At stage of the Age, an age of the synthesized speech distinguished into a baby, youth, middle age and old age.

The Speech_Rate represents a speech rate of synthesized speech.

At stage of the Length_of_Text, a length of input text sentence is represented by byte.

At stage of the TTS_Text, sentence text having optional

length is represented.

The Dur_Enable is a 1-bit flag and become a '1' when a duration time information is included in an organized data.

The FO_Contour_Enable is a 1-bit flag and become a '1' when a pitch information of each phoneme is included in the organized data.

The Energy_Contour_Enable is a 1-bit flag and become a '1' when an energy information of each phoneme is included in the organized data.

At stage of the Number_of_Phonemes, the number of phoneme needed to synthesize a sentence are represented.

At stage of the Symbol_each_phoneme, symbol such as IPA which is to represent each phoneme is represented.

The Dur_each_phoneme represents a duration time of phoneme.

At stage of the FO_contour_each_phoneme, a pitch pattern of the phoneme represented by a pitch value of beginning point, mid point and end point of the phoneme is represented.

At stage of the Energy_Contour_each_phoneme, energy pattern of the phoneme is represented and an energy value of beginning

point, mid point and end point of the phoneme is represented by decibel(dB).

The Sentence_Duration represents a total duration time of synthesized speech of the sentence.

The Position_in_Sentence represents a position of present frame in the sentence.

At stage of the offset, when the synthesized speech is interlocked with moving picture and a beginning point of the sentence is in the GOP(Group Of Pictures), a delay time consumed from beginning point of GOP to beginning point of the sentence is represented.

The Number_of_Lip_Event represents the number of changing point of lip-shape in the sentence.

The Lip_shape represents a lip-shape at lip-shape changing point of the sentence.

Text information includes a classification code for a used language and a sentence text. Prosody information includes the number of phoneme in the sentence, phoneme stream information, the duration every each phoneme, pitch pattern of phoneme, energy pattern of phoneme and is used for enhancing the natural of the synthesized speech. The synchronization information of the

moving picture with the synthesized speech can be considered as the dubbing concept and the synchronization could be realized in three ways.

Firstly, there is a method to synchronize between the moving picture and the synthesized speech by the sentence unit by which the duration of the synthesized speech is adjusted using the information about the beginning points of sentences, the durations of sentences, and the delay times of the beginning points of sentences. The beginning points of each sentence indicate locations of scenes from which output of the synthesized speech for each sentence within the moving picture is started. The durations of sentences indicate the number of scenes in which the synthesized speech for each sentence lasts. In addition, the moving picture of MPEG-2 and MPEG-4 picture compression type in which Group of Picture (GOP) concept is used should start at not any scene but a beginning scene within Group of Picture for reproduction. Therefore, the delay time of the beginning point is the information required to synchronize between the Group of Picture and the TTS and indicates a delay time between the beginning scene and a speech beginning point. This method is easy to be realized and can minimize additional effort, but is difficult to accomplish natural synchronization.

Secondly, there is a method by which beginning point information, end point information, and phoneme information are marked every each phoneme within an interval associated with

speech signal in the moving picture and these information is used to produce the synthesized speech. This method has an advantage that degree of accuracy is high since the synchronization between the moving picture and the synthesized speech by the phoneme unit can be attained but a disadvantage that additional effort should be fairly made to detect and record the duration information by the phoneme unit within the speech interval of the moving picture.

Thirdly, there is a method to record the synchronization information based on the beginning point of speech, the end point of speech, lip-shape, and a point of time of lip-shape change. Lip-shape is numeralized to distance (extent of opening) between upper lip and lower lip, distance (extent of width) between left and right and points of lip, and extent of projecting of lip and is defined as a quantized and normalized pattern depended on articulation location and articulation manner of phoneme on the basis of pattern with high discriminative property. This method is a method to raise efficiency of synchronization, while additional effort to produce the information for synchronization can be minimized.

The organized multimedia input information which is applied to the present invention allows an information provider to select and implement optionally among 3 synchronization methods as described above.

15

In addition, the organized multimedia input information is also used in the process to implement lip animation. Lip animation can be implemented by using phoneme stream prepared from the inputted text in the TTS and the duration every each phoneme, or phoneme stream distributed from the input information and the duration every each phoneme, or by using the information on lip-shape included in the inputted information.

The individual property information allows the user to change gender, age, and speech rate of the synthesized speech. Gender has male and female, and age is classified into 4, for example, 6-7 years, 18 years, 40 years, and 65 years. The change of speech rate may have 10 steps between 0.7 and 1.6 times of a standard rate. Quality of the synthesized speech can be diversified using these information.

FIG. 3 is a constructional view of the text-to-speech conversion system (TTS) according to the present invention. In FIG. 3, the TTS consists of a multimedia information input unit **10**, a data distributor by each media **11**, a standardized language processor **12**, a prosody processor **13**, a synchronization adjustor **14**, a signal processor **15**, a synthesis unit database **16**, and a picture output apparatus **17**.

The multimedia input unit **10** is configured as form of Table 1 and 2 and comprises text, prosody information, the information on synchronization with moving picture, the information on lip-

16

shape. Among these, requisite information is text, other information can be optionally provided by an information provider as optional item for enhancing the individual property and the natural and accomplishing the synchronization with the multimedia, and if needed, can be amended by a TTS user by means of a character input device (keyboard) or a mouse. These information is transmitted to the data distributor by each media **11**.

The data distributor by each media **11** receives the multimedia information of which the picture information is transmitted to the picture output apparatus **17**, text is transmitted to the language processor **12**, and the synchronization information is converted into data structure capable of utilizing in the synchronization adjustor **14** and transmitted to the synchronization adjustor **14**. If prosody information is included in the inputted multimedia information, this multimedia information is converted into data structure capable of utilizing in the signal processor **15** and then transmitted to the prosody processor **13** and the synchronization adjustor **14**. If individual property information is included in the inputted multimedia information, this multimedia information is converted into data structure capable of utilizing in the synthesis unit database **16** and the prosody processor **13** within the TTS and then transmitted to the synthesis unit database **16** and the prosody processor **13**.

The language processor **12** converts text into phoneme stream,

presumes prosody information, symbolizes this information, and then transmits the symbolized information to the prosody processor **13**. The symbol of prosody information is presumed from a boundary of the phrase and paragraph, a location of accent in word, a sentence pattern, and so on using the analysis result of syntax.

The prosody processor **13** takes the processing result of the language processor 12 and calculates a value of prosody control parameter other than prosody control parameter included in the multimedia information. Prosody control parameter includes duration pitch contour, energy contour, pause point, and pause length of phoneme. The calculated result is transmitted to the synchronization adjustor **14**.

The synchronization adjustor **14** takes the processing result of the prosody processor **13** and adjusts the duration every each phoneme in order to synchronize the result with the picture signal. The adjustment of the duration every each phoneme utilizes the synchronization information transmitted from the data distributor by each media **11**. First, lip-shape is assigned to each phoneme depended on articulation location and articulation manner of each phoneme and, on the basis of this, the assigned lip-shape is compared to lip-shape included in the synchronization information and then phoneme stream is divided into small groups by the number of lip-shape recorded in the synchronization information. Also, the duration of phoneme in

the small groups is calculated again using the duration information of lip-shape included in the synchronization information. The adjusted duration information is transmitted to the signal processor **15**, included in the processing result of the prosody processor **13**.

The signal processor **15** receives the prosody information from the multimedia distributor **11** or the processing result of the synchronization adjustor **14** to produce and output the synthesized speech using the synthesis unit database **16**.

The synthesis unit database **16** receives the individual property information from the multimedia distributor **11**, selects synthesis units adaptable to gender and age, and then transmits data required for synthesis to the signal processor **15** in response to a request from the signal processor **15**.

As can be seen from the description described above, according to the present invention, the individual property of the synthesized speech can be realized and the natural of the synthesized speech can be enhanced by organizing the individual property and prosody information presumed by the analysis of actual speech data, along with text information, as multistage information. Furthermore, a foreign movie can be dubbed in Korean by implementing the synchronization of the synthesized speech with the moving picture by way of the direct use of text information and lip-shape information which is presumed by the

19

analysis of actual speech data and lip-shape in the moving picture for the production of the synthesized speech. Still furthermore, the present invention is applicable to a variety of field such as communication service, office automation, education and so on by making the synchronization between the picture information and the TTS in the multimedia environment possible.

Although the present invention and its advantages have been described in detail, it should be understood that various changes, substitutions and alterations can be made herein without departing from the spirit and scope of the invention as defined by the appended claims.

It is therefore intended by the appended claims to cover any and all such applications, modifications, and embodiments within the scope of the present invention.

20